DOE/EIA/TR-0245

0245

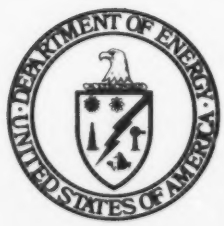
Technical Report

Procedures For Analyzing Data For Use In Models

NEWARK PUBLIC LIBRARY
NOV 19 1980
U. S. GOVT. PUBS.
DEPOSITORY 3755

June 1980

U.S. Department of Energy
Energy Information Administration
Assistant Administrator for Applied Analysis
Washington, D.C. 20585



E 3.26/3:

This publication is available from the Superintendent of Documents, U.S. Government Printing Office. An order form is enclosed for your convenience. Send order form and payment to:

Superintendent of Documents
U.S. Government Printing Office
Washington, D.C. 20402

Order Desk (202) 783-3238

Stock Number: 061-003-00150-5
Price: $ 2.00

For questions on energy statistics or information on availability of other EIA publications, contact:

U.S. Department of Energy
Energy Information Administration
Office of Energy Information Services, EI-60
Forrestal Building
Mail Stop 1F048
Washington, D.C. 20585

(202) 252-8800

Released for Printing: October 21, 1980.

DOE/EIA/TR-0245
Dist. Category UC-13

Technical Report

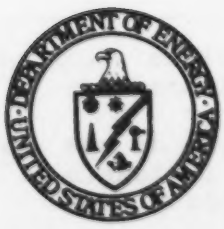
Procedures For Analyzing Data For Use In Models

June 1980

TR/OA/80-08

Prepared by:
John H. Herbert
Office of Analysis Oversight and Access

U.S. Department of Energy
Energy Information Administration
Assistant Administrator for Applied Analysis
Washington, D.C. 20585



PROCEDURES FOR ANALYZING DATA FOR USE IN MODELS

Introduction

Prior to the advent of the computer, statisticians and other analysts tended to look hard at collected data. They employed various statistical and graphical procedures in an ad hoc, nonstatistical fashion.[1] This care was in part motivated by the amount of time required to perform calculations. Since the time required was so great, the analyst was careful not to waste precious analytic time calculating and interpreting faulty data. As a result, analysts often discovered critical anomalies in a data base that specialists missed in the subject matter being examined.[2]

With the advent of increasingly efficient computers, care in the examination of data has frequently been disregarded. However, today, when energy data is employed to direct energy policy that affects the price and availability of different energy resources,and when perspectives

---

1/ Such activity is best referred to as data, rather than statistical, analysis.

2/ For an excellent example of such analysis the classical research paper by P.C. Mahalanobis is highly recommended [1].

of subject matter specialists are not broad enough or objective enough to spot anomalies in these complex data bases, there is increasing need for careful and independent analysis of data.[3]

This brief report presents an approach to energy data verification. It discusses the nature of data verification analysis and suggests some techniques for performing this type of analysis. It hopefully will encourage further discussion.

The discussion begins with a delineation of four steps for performing a data verification. The nature of data verification and the proposed methodology are then discussed at some length in general terms. A separate section examines some applications of the proposed methodology. This includes, in the context of this report, a possibly over-detailed examination of a particular diagnostic for evaluating the quality of data using regression models as a framework for analysis. This amount of attention is not meant to imply that great

---

3/ It is of interest that over the last several years there have appeared texts that contain quantitative procedures or graphical displays especially useful for data verification (see references 2,3,4,5, and 6). These sources set forth other approaches to data verification. It is also important to recognize that the term data verification is used in the present report in a restricted sense to mean that a consistency check or comparison is being made between an observation(s) and similar observation or between an observation(s) and a representation of similar observation(s).

weight is assigned the employment of such techniques since a much greater weight is assigned the choice of a model to represent the physical, financial, or economic processes that underlie the data. The applications section is augmented by illustrative examples taken from the entitlement program.

A Data Verification Procedure

Four steps are suggested to accomplish a data verification. The first step in analysis is to form models.[4] Models are viewed as building blocks that enable selective analysis of information elements to proceed. Model formulation requires specification of relationships among data elements.

The second step is to propose hypotheses of testable relationships among data elements at points in time and over time.

The third step is to select a techniques that allows the comparison of expected relationships with information actually collected.

The fourth step is to interpret the output from the selected techniques and to identify data elements not consistent with hypotheses and/or with the main body of the data. This is accomplished by means of graphical

---

4/ The sense of the term model, as used, is non-statistical.

displays and statistical analysis of the data. The objective is to identify anomalies in the data base as concretely as is possible.

By proceeding in the above outlined manner, information requiring further examination is revealed. Field investigations could provide further examination. Accordingly, field investigations can be planned in terms of information obtained from the data analysis. Field investigations are viewed as a means of determining why particular information is different from expected information. In this way, problems such as accurately filling out data forms or determining real world information on why data was so recorded might be revealed. Knowledge about the information system (data base) and its validity would, therefore, be advanced.

As stated, analysis requires proposing a variety of general models that specify possible relationships between data elements. If possible, statistical models also are proposed. Statistical models specify distributional characteristics of observations on data elements.

The specification of general or statistical models in the manner stated above is opposite to classical examinations of relationships. In classical examinations, mathematical

and statistical models are known and specified prior to collecting data or proposing tests. Interpretation of test results are, therefore, straightforward.

In data verification the output of statistical tests are considered gauges or pseudo statistics, not statistics for testing hypothesized relationships.[5] There is a need to extract information about a data set rather than confirm information about a model from a data set(s). Therefore, it is necessary to take a different tact in interpreting test results. The tact is one of exploration rather than confirmation, since it is necessary to construct models as part of data analysis rather than construct models independent of data analysis.

The fact that the data elements in any data base are ordinarily related suggests that hypothesis specification and data verification is always possible. Relatedness and hypothesis formulation could be a sum of elements in a data base being equal to another data element in the same or another data base, or it could be an increase in the value of one element being associated with

---

5/ For discussions of statistical analysis as gauges see references [2] and [8].

an increase in the value of another element, or observations on the same element over time having some pattern of consistency or randomness. The specification of all such possible expected relationships (hypothesis specification) can be considered the fundamental task of analysis.

As part of analyses strategy, it can be recommended that data elements be examined in as broad a quantitative context as seems reasonable. This includes attempts at constructing multidimensional or multivariate models. Figuratively speaking, this can be viewed as extending the quantitative perspective (or increasing the informational context) on observations in order to better gauge the relatedness of each individual observation. The reason for attempting multivariate or multidimensional models is suggestively illustrated in Figure 1.

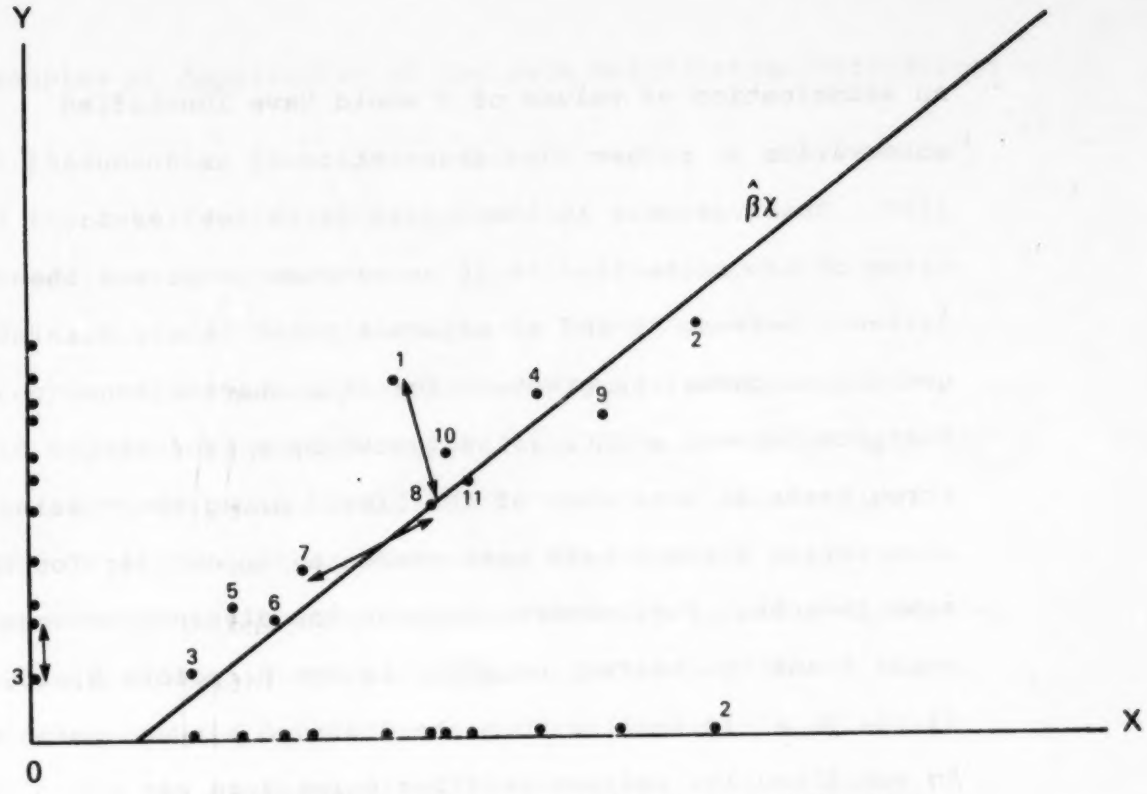


Figure 1. The Identification of Outliers

Displayed in Figure 1 is a set of data points plotted in two dimensions. A regression line has been drawn to summarize the general relationship between Y and X. Using this line as a means of comparing the different observations, it is revealed that, since the distance between an observation and the line is greatest for observation 1, this is an "outlier."[6] Moreover, an examination of values for Y and X individually would not have resulted in identification of observation 1 and a outlier.

6/ This is a loose usage of the term. A more strict usage of the term would imply that the observation is not consistent with a population it was originally assumed to be a member of.

An examaination of values of Y would have identified observation 3, rather than observation 1, as an outlier. Observation 3 is identified as an outlier in terms of two criteria. It is an extreme point and the distance between it and an adjacent point (i.e., its nearest neighbor) is greatest for this observation. Distance between points is indicated by a line with arrow heads at both ends of the line. Along the X axis, observation 2 would have been chosen as an outlier for the same reasons. Furthermore, because the distance between point 1 and its nearest neighbor in set B (points 8, 10, 11, 4, 9, 2) is smaller than the distance between point 8 in set B and its nearest neighbor point 7 in set A (points 3, 5, 6, 7) this might also suggest the existence of two clusters of points.[7] As the number of observations increased it might be worthwhile to test for such clusters.

This simple example suggests that certain facts about the observations might only be revealed when the observations are examined in higher dimensional spaces.

7/ This would be illustrated better if the observations to the right of observation 8 indicated some change in the slope of the regression line.

Examples of Application of the Data Verification Methodology

In this section some illustrative examples of application of the data verification methodology are set forth. The data elements are taken from the Economic Regulatory Administration (ERA) Entitlements Program. The program collects information from refiners on amounts and types of petroleum they refine. The variables examined include crude runs to stills (S) and crude oil receipts (R). R is defined as the volume of crude oil booked into a refinery's inventory according to specific accounting procedures consistently and historically applied by the refinery. S is defined as the total number of barrels of crude oil input to distillation units of a refinery. These variables are important from both policy and analytical points of view and will be examined in some detail.

One method of verifying entitlement information is to compare information collected on entitlement forms with like information collected on other forms. This type of verification is always recommended. Another method of using information from other forms is to infer an expected value for entitlement information from these sources and then to compare the inferred value with the value on the entitlement form. An example of such a procedure is now presented.

Data verification is begun with the specification of a model such as,

1. $R_t - S_t - L_t - DI_t \equiv 0$

where R_t = crude oil received,

S_t = runs to stills,

L_t = total losses,

I_t = inventory at end of the t th time period,

$DI_t = I_t - I_{t-1}$, the expected value of which over an extended time period n is 0,

t = time index for each refiner i.

Step two requires the formulation of hypothesis. Two possible hypothesis are:

2. $|S_t - \hat{S}_t| = 0$ or $|S_t - \hat{S}_t| \sim 0$

where S_t = runs to stills on entitlement forms

$\hat{S}_t$ = an expected value for runs to stills. (to be explained)

These hypotheses state that the expected absolute value differences between an S_t on entitlements forms and an expected $\hat{S}_t$ (e.g., from the logical model) are equal to zero,or very small.

Step three requires the specification of a technique to calculate expected values so that these values can be compared with values actually collected on entitlement forms. In the present example, this is simply accomplished by a definitional substitution and then substitution of numerical values for the data elements.

3. $R_t - S_t - L_t + I_{t-1} - I_t = 0$

where $L_t = l_{t1} + l_{t2} + l_{t3}$ for refiner i in time period t,

l_{t1} = refinery losses in time period t,

l_{t2} = crude oil used as refinery fuel in time priod t,

l_{t3} = crude oil used for petrochemical feedstock in time period t.

R, S, I as previously defined.

All informational elements identified above, except S, are obtained from forms other than entitlement information forms.[8] Therefore, the procedure is to substitute

---

8/ Often definitional problems are significant. For example, if one gets R_t from EIA-87, the definition is different from EIA-49. Whether or not these definitional differences can be adjusted is a difficult problem, however, the general approach is still valid.

values for these variables from these other forms for each time period t and compute S_t as defined below.

$$4. \quad \hat{S}_t \equiv R_t - l_{t1} - l_{t2} - l_{t3} + I_{t-1} - I_t$$

The fourth step is accomplished by examining $|S_t - \hat{S}_t|$ and if this difference is great for a large number of time periods t for a refiner, there is cause to be concerned about the accuracy of the data reported by this refiner. Another procedure might be to simply calculate the average value (i.e., $\sum_t^n |S_t - \hat{S}_t|/n$) and then to compute the ratio of the actual value for a time period $|S_t - \hat{S}_t|$ to the absolute average value, that is to calculate $|S_t - \hat{S}_t| / (\sum_t |S_t - \hat{S}_t| / n)$, and then to scrutinize those years for which the ratio is greatest. The analyst also might wish to graph the output by means of box plots [12] or some such plotting device to clearly identify and display distinctively different observations.

Before presenting another example, equation (3) is examined to reveal further details about the relationship between runs to <u>stills</u> and <u>receipts</u>. This examination will result in the formulation of several models. Equation (1) can be conveniently reexpressed as:

$$5. \quad R_t - S_t - L_t \equiv 0.$$

The term DI has been dropped as a matter of convenience and because the value of DI is expected to be zero for any time period t if data is examined that covers a sufficiently long period of time. In this equation, as L_t becomes greater and greater, S_t should become smaller and smaller in absolute value. This result follows from the satisfaction of the identity. Furthermore, for any time period t and over time, the expected value of L is not expected to be zero. Therefore, it is expected that $R_t \neq S_t$ or $R_t - S_t \neq 0$. It is also reasonable to propose that losses increase as the amounts of petroleum that are run to stills (S_t) increase. Therefore, the difference between R and S (i.e., R-S) can be expected to increase as S increases. This results in the formulation of the model presented below.

$$6. \quad B_{it} = (R_t - S_t)/S_t \quad t=1, \ldots, n$$

$$= Z_t/S_t$$

where B_{it} = indexes that are computed for each refiner i

R_t = receipts

S_t = runs to stills

$Z_t = R_t - S_t$

Having formulated the model, the next step is to propose an hypothesis. The hypothesis examined is that B is small, slightly positive, and constant. The hypothesis is formally stated as

$$8. \quad E(B_{it}) - \gamma_{it} = 0, \text{ or } E(B_{it}) = \gamma_{it} ; \; e > \gamma_{it} > 0.$$

where e = some small number

The essential feature of the above model and hypothesis is that as S increases by a certain factor (operations increase), R-S is expected to change by the same factor (losses are proportional to level of operation).

Step three in the present example would merely require calculating the difference $(R_t - S_t)$ and then dividing the difference by S_t for each time period t. This is done to compare the constancy of the ratio for different time periods for a refiner. An average value also is computed. This is accomplished by dividing $\sum_{t}^{n} B_{it}$ by n (the number of time periods). This number is used to compare the constancy of the ratio for different refiners.

The fourth step can be accomplished by showing the calculated $\bar{B}_i$ and the B_{it} in a graphical display such as a histogram. Visual examination would clearly reveal consistent and, most importantly, inconsistent (anomalous) calculated values of $\bar{B}_i$ and B_{it}.[9]

Another way of formulating the relationship between runs to stills and the difference between runs to stills and receipts is by means of a regression model. A direct result of employing this model is the calculation of certain statistics that aid in the identification and interpretation of observations not consistent with the majority of observations.

The regression model formulated is

9. $z = f(S)$

10. $Z_{it} = \hat{B}_i S_{it} + e_t$

where $\hat{B}_i$ = a constant estimated by employing some curve fitting or regression technique

S_{it} = runs to stills by firm i during some time period t

9 Plotting and graphing output is an important consideration. Many useful ways of graphing output are available and many of these have been set forth by John Tukey [3]. The choice of a graphing technique is determined by its ability to clearly display anomalies in a data set.

$Z_{it} = R_{it} - S_{it}$

R_{it} = receipts by firm i in time period t

e_t = randomly distributed residuals about the line or plane.

The hypotheses are that $\hat{B}_i$ is small and constant for all refiners and $\hat{D}_t$ (to be defined) is less than some predetermined value for all time periods for any refiner. The hypotheses are:

11. $E(B_i) - Z = 0$ or $E(B_i) = Z$, for all refiners where E indicates an expected value.

12. $\hat{D}_t < D_t$ or $\hat{D}_t - D_t \neq 0$, for all time periods t for some refiner i

where D_t is obtained from some standard statistical table such as a table based on the F- distribution and D_t is Cook's distance (to be discussed).

The explanation of this model is similar to the preceding ratio model in that proportionate changes in S are expected to result in constant proportionate changes in Z.

Step three requires the specification of a technique and criterion for computing a value for the slope. A common regression criterion is the minimization of the squared distance between the e_t and a line. Other

techniques and criterion could have been chosen for determining the slope of the line such as minimization of the absolute difference between the e_t and a line.

The final step is to interpret the output from the application of the selected technique and to identify observations not consistent with hypotheses or not consistent with the majority of observations. Within the context of a regression model framework, an interesting diagnostic statistic is Cook's Distance ($\hat{D}_t$) . This diagnostic enables the testing for a constant $\hat{B}_{(-t)}$ for all time periods for a refiner. $\hat{B}_{(-t)}$ is the $\hat{B}_i$ that is obtained by leaving out the t th observation. In an experiemental situation omitting one observation probably would result in little change in the slope of the line. Therefore, those

time periods for which the $\hat{D}_t$ calculated from the $B_{(-t)}$ are greatest require further attention.[10]

$\hat{D}_{(t)}$ was derived by means of an analogy with the F statistic. Using $\hat{D}_{(t)}$ as a criterion, step four is accomplished in a straightforward manner. Based on the F- statistic, it affords handy interpretation for the $\hat{D}_{(t)}$'s computed.[11] For example, large values of $\hat{D}_{(t)}$ can be interpreted in a similar fashion as large values for

---

[10] The diagnostic $\hat{D}_t$ is most easily interpreted in a multi-dimensional framework where the number of independent variables is greater than one. $\hat{D}_t$, however, can be expressed as the product of two interesting statistics which can be interpreted in a straightforward manner no matter the dimension of the model. One statistic is the ratio of the variance of the t th estimated value ($V(\hat{Z}_t)$) to the variance of the t th residual ($V(R_t)$). This ratio ($V(\hat{Z}_t)/V(R_t)$) measures the relative sensitivity of an estimated $\hat{B}_i$ to potential outlying values at the t th data point(13). The other statistic is the t th studentized residual which is the t th residual (e_t) divided by the standard deviation of the t th residual. This latter statistic can be compared with points from the F distribution to identify observations different from the rest in the determination of $\hat{B}_i$. Cook's distance (10) can be expressed as the product of both these statistics that is $\hat{D}_t = t_t/p \cdot V(\hat{Z}_t)/V(R_t)$ where p = number of independent variables, t_t = the studentized residual.

[11] Handy , but lacking in statistical content.

the F- statistic. Large values for particular observations indicate that they are significant in the determination of the regression line or plane. This is a problem since all observation should be assigned equal weight for such determination.

The diagnostic also can be expressed in terms of other interesting diagnostics such as the deleted residual. The deleted residual is the distance between a line or a plane and an observation when the observation is not included in the calculation of the line or plane. In addition to the statistics and diagnostics mentioned above, a graphical analysis, such as that presented in Figure 1, is highly recommended. The graphical analysis, however, should now include, at the minimum, an analysis of the deleted residual, not just the residual. This might very well give different results as to the identification of outliers.

Furthermore, whether deleted residuals are computed and compared with the residuals on the full model or studentized residuals are calcualted and compared with values from F-distribution, the point is to understand in what way individual data points are, or are not, similar to other points in the determination of some underlying relationship.

The regression analysis and diagnostics described above are more a heuristic than a statistical tool.

However, s a statistical tool it is based on a univariate, multidimensional regression model rather than a multivariate model. Such points are important since in all data analysis it is necessary to understand assumptions that must be satisfied if analysis is to be at all statistically meaningful. For example, if the bivariate correlation model, rather than a univariate regression model, were specified, it is then important to know whether observations on X an Y can be considered as being normally distributed random variables in which changes in the magnitude of the variable tend to be associated but not related in that changes in the magnitude of one determines a change in the magnitude of the other, as in the regression model (9).

If the analyst believes that correlation analysis, rather than regression analysis, is the best way to represent the relationship that exists between Z and S, then the output of a function known as the influence function (4), and its related graph might be examined in some detail. Both the function and its graph are employed to indicate the sensitivity of the computed correlation coefficient to particular observations. In this sense, they are similar to the $\hat{D}_t$.

As stated earlier, alternative forms for specifying any models should also be explored. For example, regression model 9 might also be expressed as:

$$14. \quad Z_{it} = \hat{B}_i \ln(S_{it}) + e_{it}$$

This model specifies that changes in Z become proportionately smaller as S_{it} becomes proportionately larger. Such a model might be proposed in order to examine whether refiner losses are proportionately smaller as refiner operations increase. This is not an unreasonable assumption. It implies that larger refiners are better able to control losses.

It can be observed that the number of alternative models, hypotheses, criteria, estimating techniques, interpretations, and graphical displays to consider tend to increase as the dimension of the relationships to be examined increase.

Conclusion

The need for verification of energy data was discussed. A methodology was set forth for performing data verification. Some problems and topics important for doing data verification were discussed. Examples illustrated how the methodology might be employed. Specific statistics and graphs were recommended as useful for accomplishing data verification. Both statistics and graphs help the analyst to identify anomalies in the data base.

By employing data verification techniques and methodologies, specific problem areas in data bases, such as a refiner's difficulty in initially recording accurate information on a form or discovering more fundamental accuracy problems, are unearthed. It was suggested that field investigations can be planned in terms of information revealed in the data analysis. In this way, knowledge about the information system, its breadth and validity, can be advanced. Finally, it is believed that data verification is a most useful and creative type of energy analysis that requires increased attention.

REFERENCES

1. Mahalanobis, P.C.; Majumdar, D.N.; and Rao, C.R., "Anthropometric Survey of the United Provinces, 1941: A Statistical Study," Sankhya, Volume 9, Parts 2 and 3, 1949.

2. Mosteller, F.C. and Tukey, J.W. Data Analysis and Regression. Reading, Massachusetts: Addision-Wesley, 1977.

3. Tukey, J.W., Exploratory Data Analysis. Reading, Massachusetts: Addison-Wesley, 1977.

4. Gnanadesikan, R. Methods for Statistical Data Analysis of Multivariate Observations. New York: John Wiley, 1978.

5. Barnett, V., and Lewis, T. Outliers in Statistical Data. New York: John Wiley, 1978.

6. Belsely, D.A.; Kuh, E.; and Welsch, R.E., Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley, 1980.

7. Economic Regulatory Administration, U.S. Department of Energy, FEA Entitlements Handbook. DOE-ERA-0038, 1977.

8. Gokhale, D.V., and Kullback, S. The Information in Contingency Tables. New York: Marcel Dekker, 1978.

9. Anderson, T.W. An Introduction to Multivariate Statistical Analysis. New York: John Wiley & Sons, Inc. (1958), pp. 60-78.

10. Cook, R.D., "Detection of Influential Observations in Linear Regression," Technometrics, Volume 19 (1977), pp. 15-18.

11. Dixon, W.J., et.al. BMDP-77, Berkely, California: University of California Press (1977), pp. 803-804.

12. McGill, Robert.W., Tukey, John W. and Larsen, Wayne A. "Variations of Box Plots," The American Statistician, Volume 32, No. 1 (1978), pp. 12-16.

13. Prescott, P. "An Approximate Test for Outliers in Linear Regression," Technometrics, Volume 17 (1975), pp. 129-132.

DOE F 1340.2
(2-80)

U.S. DEPARTMENT OF ENERGY

GPO SINGLE ISSUE AND ANNUAL ORDER FORM

B

(For use in ordering EIA Publications only – Read Ordering Information Section before completing form.)

SEND ORDER FORM TO: Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 20402

Enclosed is $ _____ ☐ Check

☐ Money order, or charge to my Deposit Account No.

☐☐☐☐☐☐☐ – ☐

Order No. _____

VISA

master charge

Credit Card Orders Only

Total charges $ _____ Fill in the boxes below

Credit Card No. ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

Expiration Date Month/Year ☐☐☐☐ ☐ VISA ☐ Master Charge

PLEASE PRINT OR TYPE

NAME AND ADDRESS

NAME – FIRST, LAST

COMPANY NAME OR ADDITIONAL ADDRESS LINE

STREET ADDRESS

CITY STATE ZIP CODE

(OR COUNTRY)

FOR OFFICE USE ONLY

QUANTITY	CHARGES
ENCLOSED	
TO BE MAILED	
SUBSCRIPTIONS	
POSTAGE	
FOREIGN HANDLING	
MMOB	
OPNR	
UPNS	
DISCOUNT	
REFUND	

PLEASE PRINT OR TYPE THE GPO STOCK NO. OF SINGLE ISSUES OR ANNUALS YOU WISH TO RECEIVE:

_____

_____

_____

_____

_____

_____

_____

_____

GPO 669-025

U.S. Department of Energy
Energy Information Administration
Office of Energy Information Services
Washington, DC 20585

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE, $300

FIRST – CLASS MAIL
POSTAGE & FEES PAID
U.S. DEPT. OF ENERGY

PERMIT NO. G 20

FIRST CLASS MAIL